Automated classification of demographics from face images

Jaeger, Bastian; Sleegers, Willem; Evans, Anthony

Published in:
Social and Personality Psychology Compass

DOI:
10.1111/spc3.12520

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Received: 28 February 2019 | Revised: 16 November 2019 | Accepted: 17 December 2019

DOI: 10.1111/spc3.12520

ARTICLE

WILEY

Automated classification of demographics from face images: A tutorial and validation

Bastian Jaeger | **Willem W. A. Sleegers** | **Anthony M. Evans**

Tilburg University

Correspondence
Bastian Jaeger, Department of Social Psychology, Tilburg University, Tilburg 5000 LE, The Netherlands.
Email: bxjaeger@gmail.com

Abstract

Examining disparities in social outcomes as a function of gender, age, or race has a long tradition in psychology and other social sciences. With an increasing availability of large naturalistic data sets, researchers are afforded the opportunity to study the effects of demographic characteristics with real-world data and high statistical power. However, since traditional studies rely on human raters to asses demographic characteristics, limits in participant pools can hinder researchers from analyzing large data sets. Automated procedures offer a new solution to the classification of face images. Here, we present a tutorial on how to use two face classification algorithms, Face++ and Kairos. We also test and compare their accuracy under varying conditions and provide practical recommendations for their use. Drawing on two face databases (n = 2,805 images), we find that classification accuracy is (a) relatively high, with Kairos generally outperforming Face++ (b) similar for standardized and more variable images, and (c) dependent on target demographics. For example, accuracy was lower for Hispanic and Asian (vs. Black and White) targets. In sum, we propose that automated face classification can be a useful tool for researchers interested in studying the effects of demographic characteristics in large naturalistic data sets.

Bastian Jaeger, Willem W. A. Sleegers, and Anthony M. Evans, Department of Social Psychology, Tilburg University, The Netherlands.

We thank Debbie Ma and Hannes Rosenbusch for their valuable comments.

1 | AUTOMATED CLASSIFICATION OF DEMOGRAPHICS FROM FACE IMAGES: A TUTORIAL AND VALIDATION

Exploring systematic differences in how people behave or are treated by others as a function of their gender, age, or race has a long tradition in psychology, as well as in related fields such as economics, sociology, and law. To study the effects of demographic characteristics, researchers often draw on large naturalistic data sets. For example, scholars have investigated data from game shows (Belot, Bhaskar, & van de Ven, 2010), dating websites (Feliciano, Robnett, & Komaie, 2009), criminal trials (Blair, Judd, & Chapleau, 2012), and online peer-to-peer markets (Edelman, Luca, & Svirsky, 2017). These efforts are part of the emerging field of computational social science, which uses big data to answer questions relevant to social scientists (Lazer et al., 2009). Relying on large naturalistic data sets has several advantages: It allows for precise effect size estimates and provides direct tests of how demographic variables influence real-life outcomes. While creating such data sets can be very time-intensive, researchers can often draw on preexisting shared data sets, or data sets that were created for purposes other than psychological research.

Despite the availability of large data sets, resource constraints often lead researchers to focus on a subset of the available data (e.g., Kakar, Franco, Voelz, & Wu, 2016). Since information on targets' demographic characteristics is often not available, researchers typically use human raters to code demographic information based on face images, as people are able to identify a target's gender, age, and race with very high levels of accuracy (Bruce & Young, 2012). However, the required sample of raters vastly outnumbers the typical university participant pool. For example, acquiring ratings for 100,000 images by 15 independent judges on three characteristics requires a participant pool of 22,500 individuals.[1] It would be difficult to reach this sample size, even with access to large online participant pools (Paolacci & Chandler, 2014). Automated procedures offer a new solution to the classification of face images. While automated face classification has received considerable attention in the computer science literature, social scientists have only recently begun to utilize the technology (e.g., Edelman et al., 2017; Kosinski, 2017; Rhue & Clark, 2016). Crucially, relying on an algorithm allows researchers to work with large data sets and reduces the time spent on data collection.

The remainder of this article is organized in three parts. First, we provide a short tutorial on how to use face classification algorithms, with a more detailed tutorial provided in the supporting information. Second, we assess and compare the accuracy of two algorithms in categorizing gender, age, and race based on face images. We draw on two face databases (n = 2,805 images) to test accuracy for standardized images taken under controlled conditions in the lab and for more variable images taken from the internet. Third, we discuss advantages and disadvantages of relying on face classification algorithms, we outline ethical considerations for working with naturalistic data, and we provide practical recommendations for researchers.

2 | HOW TO USE FACE CLASSIFICATION APIs

Here, we focus on two face classification algorithms: Face++ (Megvii Inc., http://www.faceplusplus.com) and Kairos (Kairos AR, Inc., https://www.kairos.com). Both can be accessed via the openly available software R (R Core Team, 2018); they can classify—among other things—a target's gender, age, and race, and they have a variety of pricing plans.

Face++ and Kairos can be accessed via their respective application programming interface (API). An API is a way of accessing the functionality of a program via another program. APIs usually have their own website where users can access their functionality. For examples, see the demo pages of both Face++ (https://www.faceplusplus.com/attributes/) and Kairos (https://www.kairos.com/demos). Another way of accessing the functionality of an API is via code: Users can instruct a computer program to perform an "API call," which consists of a communication between a client (i.e., a user's computer) and a server (i.e., the place where the API-related computations are performed). In the following section, we briefly outline the necessary steps to use face classification APIs (see Figure 1 for an overview). A detailed tutorial on how to use APIs, including code, can be found in the supporting information.

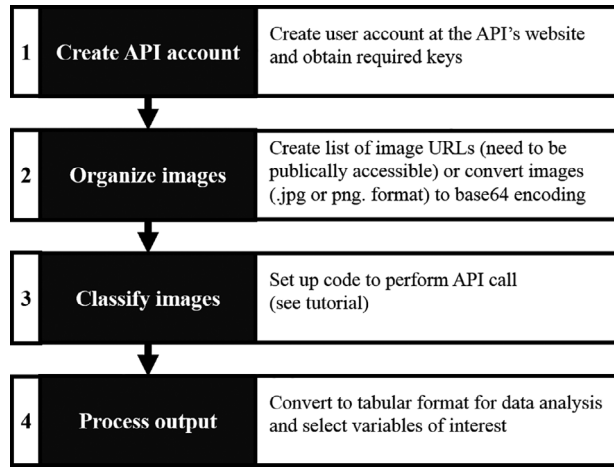

FIGURE 1 Overview of the basic steps required for using face classification application programming interfaces (APIs)

The first step is to obtain the API keys by creating an account at the website of the API classification service. Because API calls are requested computations, there is often a set of controls in place that prevent the API from being overused or abused. Typically, there is a public key (similar to a username) and a secret key (similar to a password). It is important to keep the API keys private, as others could use them and accumulate a substantial amount of processing fees. It is particularly important to remember this when sharing code, which is likely to contain personal API keys.

The second step is to organize the images to be classified. The images should either be locally stored image files or a list of URLs. Local images (i.e., images stored on your computer) should be in either a .jpg or .png file format, and URLs must refer to publically accessible images.

The third step is to perform the API call. In order to perform an API call, the public and secret API keys, the image, and the encoding of the API call itself need to be supplied. This can be accomplished in a few lines of code containing the specific information that is required by the API. The exact encoding of the API call depends on which face classification service is used. For example, Kairos requires JSON encoding and information on which face attributes to return.[2] The API documentation that can be found online indicates what is required. After performing the API call, we recommend to check whether the API call was successfully run. APIs return a status code that can be used to determine whether the call was completed successfully or whether an error occurred. If an error occurred, the status code often provides information regarding what went wrong. For example, it may be the case that the keys were incorrect, that the image file was too large, or that some of the supplied information (often referred to as arguments) was incorrectly specified or missing.

The final step is to process the returned data. Face++ and Kairos return data in JSON format. This data is organized but not necessarily suitable for data analysis. Preferably, the data is converted to tabular data, which can then be merged with other data (e.g., outcome variables or additional face attributes gathered through other means) and used for data analysis.

It may be fruitful to write code that performs the previous two steps repeatedly (i.e., a loop) to process a large amount of images. An important consideration in using such a loop is how to handle unsuccessful API calls. Unsuccessful API calls should not break the loop (thus stopping the collection of data) and should be saved to keep track of how many images could not be classified. Not all images are suitable for face classification APIs and a variety of factors, such as image quality, face size, or face rotation, may result in unsuccessful classifications. Face classification APIs differ in the extent to which they can effectively process these images of varying quality.

In the following, we present a study that tests the accuracy of the Face++ algorithm and the Kairos algorithm in classifying a target's gender, age, and race. All data, materials, and scripts are available at Open Science Framework (https://osf.io/23pn4).

3 | METHODS

3.1 | Materials

We drew on two open-access face databases, the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015) and the 10k Faces Database (Bainbridge, Isola, Blank, & Oliva, 2013), to test classification accuracy.

3.1.1 | Chicago face database

The Chicago Face Database contains images of 597 individuals taken in a controlled lab environment (Ma et al., 2015). All targets wore a gray shirt and displayed a neutral facial expression. The accompanying data set includes the self-reported gender and race of all targets. The targets' age was determined by showing each image to 20–131 (M = 43.74) participants who were asked to provide an age estimate. Age ratings were then averaged across all participants. The Chicago Face Database is particularly suited as it contains targets with widely varying demographic characteristics. Targets indicated belonging to four different racial groups (33.00% Black, 30.65% White, 18.26% Asian, and 18.09% Hispanic). Approximately half of all targets are female (51.42%), and their rated age ranges from 17 to 56 years (M = 28.86, SD = 6.30). The self-reported gender and race as well as the rated age serve as our benchmarks.

3.1.2 | 10k faces database

While the Chicago Face Database contains images of individuals varying in gender, age, and race, the images were taken under controlled conditions in the lab. However, many images people are exposed to in real life—such as profile photos on Facebook, Twitter, or Airbnb—are highly variable. To provide a more conservative test of the API's performance, we used images from the 10k Faces Database (Bainbridge et al., 2013). The full database contains more than 10,000 face images downloaded from the internet, mostly displaying nonfamous people. All images were cropped to an oval shape to eliminate background features and resized to the same height. We focus on a subset of 2,222 images for which demographic data are available. A target's gender, age, and race were determined by showing each image to 12 independent MTurk workers who categorized the faces on the relevant characteristics.[3] We excluded four targets with missing age data and two targets whose race was classified as "other." Our final data set contained 2,216 images. Targets varied in race (82.67% White, 9.93% Black, 4.15% Asian, 3.24% Hispanic) and age (11.10% younger than 20 years, 37.77% 20–30 years old, 31.68% 30–45 years old, 17.64% 45–60 years old, and 1.81% older than 65 years). There were slightly more men than women (42.69% female). The ratings provided by MTurk workers served as our benchmark.

3.2 | Procedure and analysis plan

We used the Face++ API and the Kairos API to classify the gender, age, and race of all targets. Kairos provides confidence estimates for each gender and race category, and we selected the category with the highest confidence estimate as Kairos' classification output. For each dimension, we compared the API's classification against the database-specific benchmark to determine the algorithm's accuracy. For the Chicago Face Database, the benchmark is the target's self-reported gender and race, as well as the average age estimate provided by human raters. For the 10k Faces Database, the benchmark is the gender, age, and race of targets as classified by human raters.

To estimate the performance of the APIs, we calculated their sensitivity, specificity, and accuracy. These estimates are based on the number of true positives (TP; e.g., a White individual classified as White), false positives (FP; e.g., a non-White individual classified as White), true negatives (TN; e.g., a non-White individual classified as non-White), and false negatives (FN; e.g., a White individual classified as non-White). *Sensitivity* $\left(\frac{TP}{TP+FN}\right)$ denotes the percentage of actual occurrences that were detected by the algorithm. *Specificity* $\left(\frac{TN}{TN+FP}\right)$ denotes the percentage of detected occurrences that reflect actual occurrences. *Accuracy* represents the algorithm's overall ability to discriminate between targets (e.g., accurately classifying their race) and is calculated by dividing the sum of true positives and true negatives by the total number of targets: $\frac{TP+TN}{TP+TN+FP+FN}$.

4 | RESULTS

Before analyzing the classification accuracy of the algorithms, we tested if the algorithms were able to detect a face and thus provide a classification for every image. Both Face++ and Kairos detected a face in all 597 images of the Chicago Face Database. For the more variable images of the 10k Faces Database, Face++ detected a face in all 2,216 images while Kairos detected a face in 2,208 images (99.64%). Thus, the face detection rate of both algorithms was close to 100%. The results reported here are based on all images for which both algorithm were able to provide a classification.

4.1 | Gender

To test accuracy in gender classification, we first compared the gender the Face++ algorithm assigned to a given target with the benchmark gender of the targets from both databases (see Table 1). Accuracy was at 88.94% for the Chicago Face Database, 95% confidence interval (CI) [86.15%, 91.35%] and at 90.17% for the 10k Faces Database, 95% CI [88.85%, 91.38%]. Accuracy levels did not significantly differ between the two samples, $\chi^2(1) = 0.65$, $p = .42$, $\Delta = 1.23\%$. Thus, we did not find any evidence that the performance of the Face++ algorithm in classifying gender was lower for the more variable image set.

Next, we compared the gender the Kairos algorithm assigned to a given target with the benchmark gender of the targets from both databases (see Table 1). Accuracy was at 96.15% for the Chicago Face Database, 95% CI [94.28%, 97.54%] and at 98.55% for the 10k Faces Database, 95% CI [97.96%, 99.01%]. Performance was slightly better for the more variable image set, $\chi^2(1) = 12.86$, $p < .001$, $\Delta = 2.40\%$.

TABLE 1 Accuracy of the Face++ algorithm in classifying the gender of targets from the Chicago Face Database and the 10k Faces Database

	Chicago Face Database		10k Faces Database	
	Female	Male	Female	Male
Sensitivity				
Face++	82.08%	96.21%	88.41%	91.50%
Kairos	93.16%	99.31%	98.10%	98.89%
Specificity				
Face++	96.21%	82.08%	91.50%	88.41%
Kairos	99.31%	93.16%	98.89%	98.10%
Accuracy				
Face++	88.94% [86.15%, 91.35%]		90.17% [88.85%, 91.38%]	
Kairos	96.15% [94.28%, 97.54%]		98.55% [97.96%, 99.01%]	

Finally, we compared the performance of the two algorithms. The Kairos algorithm was more accurate than the Face++ algorithm when classifying faces from both the Chicago Face Database (7.21 percentage points difference, $\chi^2(1) = 22.46$, $p < .001$) and the 10k Faces Database (8.38% difference, $\chi^2(1) = 144.11$, $p < .001$). In sum, the Kairos algorithm showed better performance in gender classification for both controlled and more variable face images.

4.2 | Age

To test accuracy in age classification, we first compared the age the Face++ algorithm assigned to a given target with the benchmark age of the targets from both databases (i.e., the error in age estimation). For the Chicago Face Database targets, the average error for estimated age was 7.98 years ($SD = 5.67$), which is significantly different from zero, $t(596) = 34.38$, $p < .001$ (Figure 2a). For the 10k Faces Database targets, the average age estimated by Face++ shifted upwards with each age category (see Figure 3a). We calculated the percentage of age estimates that fell within the benchmark age category. Across the five age categories, only 18.34% of age estimates fell within the benchmark age range. Examining the distance between targets' assigned age category and their benchmark age category showed that for the majority of targets, age estimates were only off by one category ($M = 1.11$, $SD = 0.73$).

Next, we compared the age the Kairos algorithm assigned to a given target with the benchmark age of the targets from both databases. For the Chicago Face Database targets, the average error for estimated age was 3.30 years ($SD = 2.64$), which is significantly different from zero, $t(596) = 30.58$, $p < .001$ (Figure 2b). For the 10k Faces Database targets, the average age estimated by Kairos shifted upwards with each age category (see Figure 3b). We calculated the percentage of age estimates that fell within the benchmark age category. Across the five age categories, 38.95% of age estimates fell within the benchmark age category. Examining the distance between targets' benchmark age category and their assigned age category showed that for the majority of targets, age estimates were only off by one category ($M = 0.65$, $SD = 0.56$).

Finally, we compared the performance of the two algorithms. For the Chicago Face Database targets, the Kairos algorithm was significantly more accurate than the Face++ algorithm, with an average difference in error for estimated age of 4.68 years, $t(842.46) = 18.28$, $p < .001$. For the 10k Faces Database, the majority of age estimates of both algorithms fell outside of the benchmark age category (Face++: 81.66%; Kairos: 61.05%). However, age estimates of the Kairos algorithm were significantly more often within this age range, $\chi^2(1) = 228.42$, $p < .001$, $\Delta = 20.61\%$. Moreover, the mean distance between a target's estimated age category and their benchmark age category was smaller for the Kairos algorithm, $t(4,124.1) = 23.59$, $p < .001$, $\Delta = 0.46$. In sum, for both data sets, age estimates by Kairos were more accurate than age estimates by Face++.

4.3 | Race

To test accuracy in race classification, we first compared the race the Face++ algorithm assigned to a given target with the benchmark race of the targets from both databases. Accuracy was at 72.86%, 95% CI [69.11%, 76.39%] for the Chicago Face Database and at 82.79%, 95% CI [81.15%, 84.34%] for the 10k Faces Database (Table 2). There was a significant difference in accuracy levels between the two samples, $\chi^2(1) = 29.09$, $p < .001$, $\Delta = 9.93\%$, showing that was better for the 10k Faces Database.[4]

Next, we compared the race the Kairos algorithm assigned to a given target with the benchmark race of the targets from both databases. Accuracy was at 89.28%, 95% CI [86.52%, 91.65%] for the Chicago Face Database and at 95.06%, 95% CI [94.08%, 95.93%] for the 10k Faces Database (Table 3). Accuracy levels differed significantly between the two samples, $\chi^2(1) = 26.13$, $p < .001$, $\Delta = 5.78\%$, showing that performance was better for the 10k Faces Database.

TABLE 2 Accuracy of the Face++ algorithm in classifying the race of targets from the Chicago Face Database and the 10k Faces Database

	Asian	Black	Hispanic	White
Sensitivity				
Chicago	90.83%	90.86%	-	85.79%
10k	64.13%	75.91%	-	87.83%
Specificity				
Chicago	86.27%	92.25%	-	84.54%
10k	92.68%	94.12%	-	71.88%
Accuracy				
Chicago	72.86% [69.11%, 76.39%]			
10k	82.79% [81.15%, 84.34%]			

TABLE 3 Accuracy of the Kairos algorithm in classifying the race of targets from the Chicago Face Database and the 10k Faces Database

	Asian	Black	Hispanic	Other	White
Sensitivity					
Chicago	93.58%	94.42%	66.67%	-	94.54%
10k	73.91%	95.00%	59.72%	-	97.53%
Specificity					
Chicago	98.36%	98.75%	96.29%	-	93.38%
10k	99.34%	99.65%	97.47%	-	93.49%
Accuracy					
Chicago	89.28% [86.52%, 91.65%]				
10k	95.06% [94.08%, 95.93%]				

Finally, we compared the performance of the two algorithms for both databases. Results showed that the Kairos algorithm outperformed the Face++ algorithm by 16.42 percentage points for the Chicago Face Database, $\chi^2(1) = 51.38$, $p < .001$, and by 12.27 percentage points for the 10k Faces Database, $\chi^2(1) = 167.52$, $p < .001$.

5 | GENERAL DISCUSSION

Many important social outcomes are shaped by gender, age, and race. Exploring the influence of demographic characteristics has been a topic of intense study in psychology and other social sciences. With more social interactions moving to online environments where profile photos are prevalent (e.g., economic exchange, dating, and social networking), new methods for data extraction (Landers, Brusso, Cavanaugh, & Collmus, 2016), and a general increase in the availability of data relevant for social scientists (Chen & Wojcik, 2016; Kosinski, Wang, Lakkaraju, & Leskovec, 2016; Lazer et al., 2009), researchers are afforded the opportunity to work with large naturalistic data sets. Given these developments, automated face classification can be a useful tool. We presented a tutorial and R code on how to use two face classification algorithms and tested their performance by drawing on two face databases (n = 2,805 images).

5.1 | Evaluating and comparing the algorithms' performance

Kairos correctly classified the gender of approximately 98% of targets and the race of 94% of targets. Face++'s performance was slightly lower, with 90% correct gender classifications and 80% correct race classifications. Lower

performance in race classification was partly due to the fact that Face++ does not detect Hispanic targets and all Hispanic targets in our data sets were consequently misclassified. Accuracy improved to 86% when restricting analyses to non-Hispanic targets, which was still below the accuracy level of Kairos. Classification accuracy of both algorithms varied depending on the race of the target. For example, Kairos correctly classified 98% of all White targets from the 10k Faces database but only 60% of all Hispanic targets. Face++ correctly classified 88% of all White targets but only 64% of all Asian targets. Overall, these results show that gender and race classifications by algorithms can be as accurate as classifications by human raters, whose classification accuracy is usually above 90% (Bruce & Young, 2012; Hill, Bruce, & Akamatsu, 1995; Levin & Angelone, 2002). Finally, Kairos performed better in age classification. With estimates by human raters as the benchmark, both algorithms tended to overestimate targets' age.[5] However, estimates by Kairos (mean absolute error of 3.30 years for the Chicago Face Database) were significantly closer to our benchmark than estimates by Face++ (mean absolute error of 7.98 years).

Neither algorithm showed worse performance on any characteristic when face images were not taken in highly standardized conditions but were more variable regarding image quality, lighting condition, head pose, and facial expression.[6] This observation is important as many data sets of interest contain variable photos, such as profile photos on Airbnb (Edelman et al., 2017) or screenshots of TV game show footage (Darai & Grätz, 2013). Taken together, our findings demonstrate that algorithms can provide accurate classifications of demographic characteristics, even for variable, nonstandardized images downloaded from the internet.

5.2 | Advantages and limitations of using face classification APIs

Relying on automated face classification procedures rather than human participants has several key advantages. With automated classification, a researcher's sample size is not limited by the size of their participant pool and, to a much lesser extent, by their research budget. This means that hypotheses can be tested using large sample sizes, providing high statistical power. By definition, studies with high statistical power will detect true relationships more often, thus reducing the number of false negatives in the literature. Research lines with high statistical power also tend to produce more accurate effect size estimates and a higher proportion of statistically significant results that actually reflect true relationships (Button et al., 2013; Ioannidis, 2005). In sum, high statistical power is essential for producing reliable research and recent large-scale failures to replicate established findings in psychology have led to an increased focus on power (Fraley & Vazire, 2014; Open Science Collaboration, 2015).

We also hope that the availability of easily accessible APIs will encourage researchers to test their hypotheses using large, naturalistic data sets. While studies from both the lab and the field are needed to convincingly demonstrate an effect, scholars have noted that the latter is often neglected by psychologists, calling for more analyses of real-world data (Baumeister, Vohs, & Funder, 2007; Maner, 2016). This call coincides with an increasing availability of large data sets that can be used to test psychological theories (Chen & Wojcik, 2016; Kosinski et al., 2016).

Relying on commercial software also has potential drawbacks. Our results show that the accuracy of the algorithms substantially varies across demographic groups, which means that for some data sets (e.g., samples that include a lot of Asian or Hispanic targets), overall accuracy might be relatively low. It is often unclear how algorithms operate and what data sets they were trained on. Therefore, it is crucial to rigorously test and validate algorithms before they are used in research. We provided first evidence for their validity here, but future studies need to test the algorithms under different conditions. For example, while we tested the algorithms' accuracy in classifying variable images taken from the internet, future studies should look at accuracy levels for profile photos from Facebook or Airbnb, which have been used in recent research (Edelman et al., 2017; Jaeger, Sleegers, Evans, Stel, & van Beest, 2018; Kosinski, 2017). Future studies should also test (a) how accurately the algorithms can classify a wider range of race categories, including biracial individuals, and (b) how the algorithms' performance compares to the performance of human raters. For this purpose, future studies could compare the accuracy of human and algorithmic classifications of images from face databases which include self-report data on targets' gender, age, and race (e.g., the Facelab

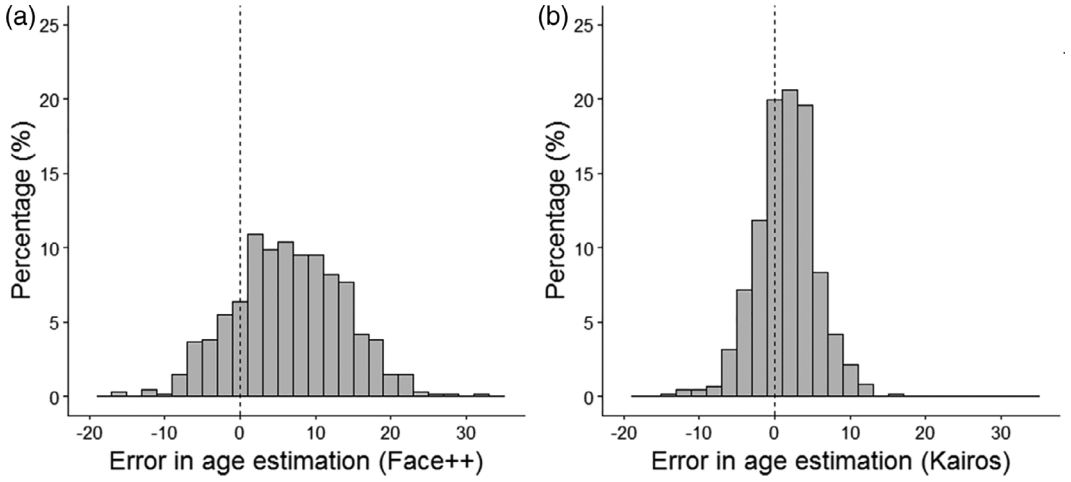

FIGURE 2 Distribution of the difference between the age estimated by (a) the Face++ algorithm or (b) the Kairos algorithm and the average age estimate of human raters for the Chicago Face Database. The dashed line represents no difference between the algorithm and human raters. Observations left of the dashed line represent an underestimation by the algorithm whereas observations right of the dashed line represent an overestimation by the algorithm

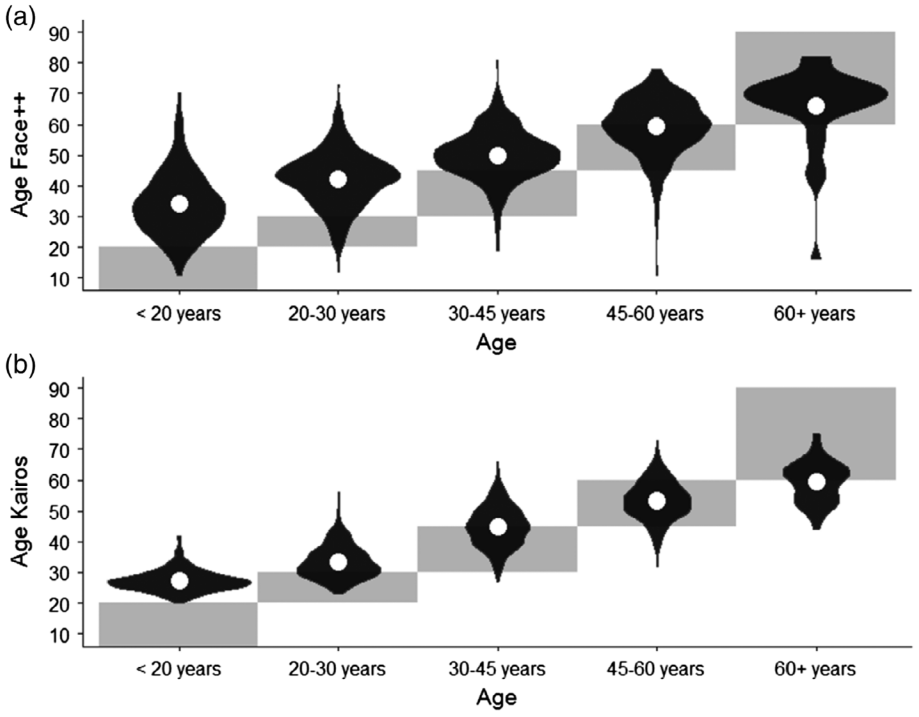

FIGURE 3 The distribution of estimated age by (a) Face++ or (b) Kairos as a function of benchmark age category of the 10k Faces Database. White dots denote the average estimated age of the algorithm. The shaded areas illustrate the targets' benchmark age categories. The overlap between the age distribution and the shaded area represents the proportion of age estimates by the algorithm that fell within the targets' benchmark age categories

London Database; DeBruine & Jones, 2017). This approach would also address a limitation of the present study in which the APIs' classifications of some dimensions were compared against averaged ratings by humans rather than self-reports. Finally, going beyond the classification of demographic characteristics, future studies could test the algorithms' performance in classifying other dimensions, such as emotion expressions or attractiveness, which are both provided by Face++.

5.3 | Ethical considerations

Conducting studies with naturalistic data sets—a context in which face classification algorithms are particularly useful—presents unique challenges to researchers who have to ensure that ethical standards are met. At the moment, there is no comprehensive set of guidelines determining when and how online data can be ethically used, and standards may vary between different institutional review boards (IRB; Chen & Wojcik, 2016; Michal Kosinski, Matz, Gosling, Popov, & Stillwell, 2015). However, this should not be taken as an excuse to dismiss ethical considerations altogether.

Many studies in computational social science rely on data that is publicly available, but that was not created for research purposes (e.g., ebay listings and social network activity). This can make it difficult or even impossible to obtain informed consent from individuals providing the data. Some have argued that public data on the internet should be treated as archival data, which can be used without informed consent (Kosinski et al., 2015). Given the lack of clear guidelines, researchers can ask themselves how likely it is that people would object to the use of their data. While a researcher's evaluation might not be objective or unbiased, there are differences in data sensitivity that most people would probably agree on. For example, the price of an item in a peer-to-peer market is easily accessible to a large audience and widely disseminating this information is often the central aim of the website's user. Other types of data, such as sexual preferences disclosed on a dating website, are more sensitive and people might be more likely to object to their use for research purposes. If a study deals with such data, attempts could be made to obtain informed consent from the relevant individuals.

A related issue is the anonymization of data. Researchers need to ensure that any identifying information is removed when data sets are shared. In some contexts, this might be more difficult than anticipated. While it is relatively easy to remove obvious identifiers such as names or IP addresses, a person's identity can often be inferred from other information. For example, in the context of Airbnb, it might be possible to identify a host from a combination of data points such as the neighborhood they live in, the size of their apartment, and the price they are asking. Guaranteeing a person's anonymity is a particularly important issue when dealing with personal photos. Just like other personal identifiers, photos should not be shared without the person's consent. Here, relying on an algorithm to classify images can actually help ensure anonymity as the images do not have to be shown to human participants in order to collect demographic information.

In sum, researchers should to be aware that ensuring ethical standards is particularly challenging when dealing with large sets of naturalistic data that individuals did not provide for research purposes. Even if there are no clear restrictions regarding the use of a specific data set, researchers should consult their IRB to ensure that broader ethical guidelines are met.

5.4 | Practical recommendations

There are several ways in which the use of face classification algorithms can be optimized. For gender and race classification, Kairos provides confidence estimates for each category. Here, we selected the category with the highest confidence estimate as the detected category. However, researchers can also exclude images that could not be classified with a predetermined level of confidence. For example, when studying the effect of race in a large data set,

researchers could restrict their analyses to images for which the algorithm was able to determine the target's race with at least 90% confidence. Excluding images will lower sample size, but this might be a price worth paying to reduce error in classifications, especially when the initial data set is large. At the same time, researchers need to be aware that systematic exclusion of images might introduce selection bias. For example, setting a high confidence threshold for race classifications might lead to more accurate classifications, but also to a disproportionately high exclusion rate of Hispanic targets for whom classification accuracy is lower. Given a large enough sample, we recommend that the robustness of any effect is tested by varying the confidence threshold for classifications. In general, researchers should be aware of the characteristics of their image set. As our results have shown, classification accuracy is dependent on several factors. Researchers need to manually examine at least a part of their image set to check whether image properties allow for accurate classifications.

6 | CONCLUSION

Large naturalistic data sets afford researchers to test their theories with high statistical power using data that reflects real-world behavior. For researchers studying the influence of demographic characteristics, this can be a challenge since a large number of participants is needed to classify targets' gender, age, or race. The results presented here suggest that algorithms can provide relatively accurate classifications of demographic characteristics. In some (but not all) aspects, their performance is close to the performance of human raters. Face classification algorithms are easy to use and more time-efficient, therefore providing a useful alternative to human raters.

ORCID

Bastian Jaeger https://orcid.org/0000-0002-4398-9731

ENDNOTES

[1] This calculation assumes that each participant takes 20 min to rate a total of 200 images on one characteristic.

[2] JSON is a language-independent format for transmitting and receiving information.

[3] A target's age was determined by taking the average estimated age across the 12 raters. Targets were then categorized into one of five age groups. A target's gender and race were determined by taking the modal response of raters.

[4] Face++ does not provide a classification for Hispanics. Accuracy was at 88.96%, 95% CI [85.84%, 91.59%] for the Chicago Face Database and at 85.58%, 95% CI [84.02, 87.04] for the 10k Faces Database when we focused only on non-Hispanic targets.

[5] Previous studies have shown that age estimates by human raters are relatively accurate, with mean absolute differences between actual age and estimated age of around 5 years (George & Hole, 2000; Han, Otto, Liu, & Jain, 2015; Sörqvist & Eriksson, 2007; Voelkle, Ebner, Lindenberger, & Riediger, 2012).

[6] It should be noted that even though some aspects of the face images were variable, all faces took up a large part of the image and were approximately photographed from the front. More research is needed to determine how specific aspects of the images affect the algorithms' performance.

Further Reading

Jaeger, B., Sleegers, W. W. A., Evans, A. M., Stel, M., & van Beest, I. (2019). The effects of facial attractiveness and trustworthiness in online peer-to-peer markets. *Journal of Economic Psychology*, *75*, 102125.

REFERENCES

Bainbridge, W. A., Isola, P., Blank, I., & Oliva, A. (2013). Establishing a database for studying human face photograph memory. In N. Miyake, D. Peebles, & R. P. Coopers (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 1302–1307). Austin, TX: Cognitive Science Society.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Belot, M., Bhaskar, V., & van de Ven, J. (2010). Promises and cooperation: Evidence from a TV game show. *Journal of Economic Behavior and Organization*, *73*(3), 396–405. https://doi.org/10.1016/j.jebo.2010.01.001

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2012). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science*, *15*(10), 674–679.

Bruce, V., & Young, A. (2012). *Face perception. London*. New York: Psychology Press.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, *21*(4), 458–474. https://doi.org/https://doi.org/10.1037/met0000111

Darai, D., & Grätz, S. (2013). Attraction and cooperative behavior. Retrieved from http://www.neweconomists.org/files/Attraction.pdf

DeBruine, L. M., & Jones, B. C. (2017). *Face research lab London set*. figshare. Retrieved from https://doi.org/10.6084/m9.figshare.5047666.v3

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, *9*(2), 1–22. https://doi.org/10.1257/app.20160213

Feliciano, C., Robnett, B., & Komaie, G. (2009). Gendered racial exclusion among white internet daters. *Social Science Research*, *38*(1), 39–54. https://doi.org/10.1016/j.ssresearch.2008.09.004

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One*, *9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

George, P. A., & Hole, G. J. (2000). The role of spatial and surface cues in the age-processing of unfamiliar faces. *Visual Cognition*, *7*(4), 485–509. https://doi.org/10.1080/135062800394621

Han, H., Otto, C., Liu, X., & Jain, A. K. (2015). Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(6), 1148–1161. https://doi.org/10.1109/TPAMI.2014.2362759

Hill, H., Bruce, V., & Akamatsu, S. (1995). Perceiving the sex and race of faces: The role of shape and colour. *Proceedings of the Royal Society B: Biological Sciences*, *261*(1362), 367–373. https://doi.org/10.1098/rspb.1995.0161

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), 0696–0701. https://doi.org/10.1371/journal.pmed.0020124

Kakar, V., Franco, J., Voelz, J., & Wu, J. (2016). The visible host: Does race guide Airbnb rental rates in San Francisco? Retrieved from https://mpra.ub.uni-muenchen.de/78275/

Kosinski, M. (2017). Facial width does not predict self-reported behavioral tendencies. *Psychological Science*, *28*(11), 1675–1682. https://doi.org/10.1177/0956797617716929

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences. *American Psychologist*, *70*(6), 543–556. https://doi.org/10.1037/a0039210

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, *21*(4), 493–506. https://doi.org/10.1037/met0000105

Landers, R. N., Brusso, R., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods*, *21*(4), 475–492. https://doi.org/10.1037/a0033269

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., … Van Alstyne, M. (2009). Life in the network: The coming age of computational social science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742.Life

Levin, D. T., & Angelone, B. L. (2002). Categorical perception of race. *Perception*, *31*(5), 567–578. https://doi.org/10.1068/p3315

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Maner, J. K. (2016). Into the wild: Field research can increase both replicability and real-world impact. *Journal of Experimental Social Psychology*, *66*, 100–106. https://doi.org/10.1016/j.jesp.2015.09.018

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188. https://doi.org/10.1177/0963721414531598

R Core Team. (2018). R: A language and environment for statistical *Computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.r-project.org/

Rhue, L., & Clark, J. (2016). *Who gets started on Kickstarter? Racial disparities in crowdfunding success*. Retrieved from http://www.ssrn.com/abstract=2837042

Sörqvist, P., & Eriksson, M. (2007). Effects of training on age estimation. *Applied Cognitive Psychology*, *21*(1), 131–135. https://doi.org/10.1002/acp.1271

Voelkle, M. C., Ebner, N. C., Lindenberger, U., & Riediger, M. (2012). Let me guess how old you are: Effects of age, gender, and facial expression on perceptions of age. *Psychology and Aging*, *27*(2), 265–277. https://doi.org/10.1037/a0025065

AUTHOR BIOGRAPHIES

Bastian Jaeger is an Assistant Professor at Tilburg University (Department of Social Psychology) in the Netherlands. His work primarily focuses on face perception and social decision making. Specifically, he investigates how people infer different characteristics from facial features, how accurate these inferences are, and how they affect decision making. He is also interested in exploring new methods to study social psychological phenomena.

Willem Sleegers is an Assistant Professor at Tilburg University (Department of Social Psychology) in the Netherlands. His research primarily focuses on beliefs and meaning. Specifically, he investigates how people form and update beliefs, how we can change their beliefs, and what the role of physiological arousal is in responding to belief feedback.

Anthony Evans is an Assistant Professor at Tilburg University (Department of Social Psychology) in the Netherlands. His research primarily focuses on trust and cooperation. Specifically, he investigates how people decide whom to trust, the cognitive processes underlying trust decisions, and the role of trust in online marketplaces.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Jaeger B, Sleegers WWA, Evans AM. Automated classification of demographics from face images: A tutorial and validation. *Soc Personal Psychol Compass*. 2020;14:e12520. https://doi.org/10.1111/spc3.12520